DOCUMENT RESUME

ED 149 619 FL 009 206

AUTHOR Massaro, Dominic W.
TITLE An Information-Processing Model of Understanding Speech.
PUB DATE Sep 77
NOTE 37p.; Paper presented at the NATO Symposium on Language Interpretation and Communication (Venice, Italy, September, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Acoustics; Audition (Physiology); Auditory Perception; Aural Stimuli; *Cognitive Processes; Communication (Thought Transfer); Language Research; Listening; *Listening Comprehension; Memory; Models; Psychoacoustics; *Psycholinguistics; *Speech Communication; *Verbal Communication
IDENTIFIERS *Language Processing

ABSTRACT

Language processing is the abstraction of meaning from a physical signal, such as a sequence of speech sounds. Processing a spoken message requires a series of transformations that begin with the acoustic signal arriving at the ears and end with meaning in the mind of the listener. The goal of this information-processing model is to describe how speech is processed, not simply what the listener must know to understand speech. Language processing is viewed as a sequence of internal processing stages or operations that occur between stimulus and meaning. The operations of a particular stage take time and transform the information in some way, making the transformed information available to the following stage of processing. Two theoretical components or constructs are important in the model. First, the structural component describes the nature of the information at a particular stage of processing. Second, the functional component describes the operations of a stage of processing. The information-processing model is used heuristically to incorporate data and theory from a variety of studies of speech processing. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED149619

An Information-processing Model of Understanding Speech*

Dominic W. Massaro

University of Wisconsin

Madison, Wisconsin 53706 U.S.A.

1. Introduction

2. Feature Detection

   a. Acoustic Features

   b. Acoustic Features in Fluent Speech

3. Primary Recognition

   a. Preperceptual Auditory Storage

4. Secondary Recognition

   a. Perceptual and Contextual Contributions

5. Rehearsal and Recoding

   a. Generated Abstract Memory

   b. Rehearsal

   c. Recoding

6. Long-Term Memory

7. Conclusion

* Paper to be presented at the N.A.T.O. Symposium: Language Interpretation and Communication, Venice, Italy, September, 1977.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Dominic W. Massaro

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

FL009206

Abstract

Language processing is the abstraction of meaning from a physical signal such as a sequence of speech sounds. Processing a spoken message requires a series of transformations that begin with the acoustic signal arriving at the ears and end with meaning in the mind of the listener. The goal of this information-processing model is to describe how speech is processed, not simply what the listener must know to understand speech. Language processing is viewed as a sequence of internal processing stages or operations that occur between stimulus and meaning. The operations of a particular stage take time and transform the information in some way, making the transformed information available to the following stage of processing. Two theoretical components or constructs are important in the model. First, the structural component describes the nature of the information at a particular stage of processing. Second, the functional component describes the operations of a stage of processing. The information-processing model is used heuristically to incorporate data and theory from a variety of studies of speech processing.

1. Introduction

Simultaneous interpretation provides an ideal situation for a logical analysis of an information-processing model. To the external observer, the interpreter listens to a speech in one language, while simultaneously (or at least within a few seconds or so) articulates the same message in a second language. In terms of an information-processing description, the simultaneous interpreter must decode the surface structure of the original message, map it into some abstract representation, take this same abstract representation and map it into a new surface structure, and finally articulate the translated message.

Almost every aspect of information-processing research and theory is relevant to language interpretation. The initial decoding and analysis of the spoken message by the listener is a pattern-recognition problem. Work in speech perception research should contribute to our description of this initial stage of language interpretation. Understanding the decoded message goes hand in hand with its perception, but additional processes are critical. If the listener did not "know" the language being spoken, many of the patterns of syllables might be recognized, but very little meaning could be imposed. To know the language means to have knowledge of the correspondences between the perceptual representations of the language

and the conceptual ones. Psychologists and psycholinguists have recently developed a plethora of descriptions of the representation and utilization of conceptual knowledge (e.g., Kintsch & Schank, this volume). It is at this stage that language appears magical, but probably not unique, since one could argue for similar underlying representations in music experience (for example, Bernstein, 1976). Once the interpreter obtains meaning at some level, he can actually translate the message by mapping the meaning onto the new surface structure. This skill would seem to reduce to one that even the unilingual has; for example, he/she might be asked to describe a recent book that was read, or what did John say at the party last night.

Our analysis of the language interpretation and communication situation would seem to imply that no unique or novel skills are required, as long as the interpreter knows the two relevant languages as well as the person on the street knows one. This analysis must be wrong, however, or else why would this august group be assembled on the island of San Giorgio? The interpreter is clearly more than the simple sum of the component skills that we have discussed. Time-sharing between processes and simultaneous processing are necessary for language interpretation. Deriving the deep structure message and selecting the appropriate surface form must go on simultaneously with decoding the speech currently being heard. But our person communicating in San Marco must also time-share and parallel-process in the unbroken chain of political discussion, although he/she

operates in terms of just one surface form whereas our interpreter must work in two. The apparent similarities in language interpretation and everyday language usage would seem to imply that what we know about normal language usage would be relevant to understanding simultaneous translation. This paper reviews the current state of the art in understanding spoken language, in terms of a general information-processing model. It is hoped that the researcher interested in simultaneous interpretation will be able to encode the surface message into something of value in this regard.

The present paper describes some of the processing stages involved in language performance along with relevant research questions. We use an information-processing model as a heuristic device to analyze theory and data from a variety of approaches. Language performance begins with the language stimulus and involves a sequence of internal processing stages before communication occurs. The processing stages are logically successive although they do overlap in time. Each stage of information processing operates on the information that is available to it and makes this transformed information available to the next stage of processing.

The speech stimulus consists of changes in the atmospheric pressure at the ear of the listener. The listener is able to experience the continuous changes in pressure as a set of discrete percepts and meanings. Our goal is to analyze the series of processing stages that allow this impressive transformation to take place. Figure 1 presents a flow diagram of the temporal course of perception of a language pattern such as speech. At each stage the

system contains structural and functional components. The structural component represents the information available at a particular stage of processing. The functional component specifies the procedures and processes that operate on the information held in the corresponding structural component. The model distinguishes four functional components; feature detection, primary recognition, secondary recognition, and rehearsal-recoding. The corresponding structural components represent the information available to each of these stages of processing. The stages will now be described in more detail, along with some theoretical issues and relevant research.

2. Feature Detection

The changes in sound pressure set the eardrums in motion and these mechanical vibrations are transduced into a set of neural impulses. The neural impulses have a direct relationship to the changes in mechanical vibrations. We call the transformation from mechanical to neural information, feature detection and evaluation. The signal in the form of continuous changes in vibration pattern is transformed into a set of relatively discrete features. Features do not have to be relatively primitive such as the amount of energy in a particular frequency band, but they may include information about the duration or rate of intensity and frequency change. It would be possible, for example, to have a feature detector that responds to the rising first format transition that is characteristic of the class of stop consonants.

a. Acoustic features

One traditional concern in speech research has been to determine the acoustic features that are utilized in perception. In terms of our model, the feature detection process places features in a brief temporary storage called preperceptual auditory storage (PAS), which holds information from the feature detection process for about 250 msec. The primary recognition process integrates these features into a synthesized percept which is placed in synthesized auditory memory. One critical question is what features are utilized and a second important question is how are all of the features integrated together. Does the listener only process the least ambiguous feature and ignore all others, or are the features given equal weight, and so on? Despite the overwhelming amount of research on acoustic features, very little is known about how the listener puts together the mulititude of acoustic features in the signal in order to arrive at a synthesized percept.

The integration of acoustic features has not been extensively studied for two apparent reasons. The first is that research in this area was highly influenced by linguistic descriptions of binary all-or-none distinctive features (Jakobson, Fant, & Halle, 1961). One of the goals of distinctive feature theory was to minimize the number of distinctive features of the language. Therefore, distinctive features were designed to be general and not stimulus specific. If a distinctive-feature difference distinguished two

phonemes in the language, that same distinction was assumed to distinguish several other phoneme pairs. Given the distinctive feature of voicing, then, the distinction of voiced versus voiceless can account for the differences between /z/ and /s/, /v/ and /f/, and so on. Given the assumption of a binary representation of distinctive features the integration of information from two or more dimensions would be a trivial problem. Integrating binary features from voicing and place of articulation, for example, could be carried out by simple logical conjunction. If the consonant /b/ were represented as voiced and labial and /p/ were represented as voiceless and labial, the identification of voiced labial sound would /b/ whereas the identification of a voiceless labial sound would be /p/.

A second reason for the neglect of the integration problem is methodological. The primary method of study involved experiments in which the speech sound was varied along a single relevant dimension. For example, in a study of voicing all voicing cues were made neutral except one, such as voice onset time and then this dimension was varied through the relevant values. Similarly, place of articulation was studied by neutralizing all cues but one, and then varying the remaining dimension through the appropriate values. Very few experiments independently varied both voicing cues and place cues within a particular experiment. Therefore, little

information was available about how these cues were integrated into a synthesized percept.

More recently, we have initiated a series of experiments that are aimed more directly at the study of the integration of acoustic features in speech perception (Massaro & Cohen, 1976; Oden & Massaro, 1977). In contrast to the traditional linguistic description, we assume that the acoustic features held in preperceptual auditory storage (PAS) are continuous, so that a feature indicates the degree to which the quality is present in the speech sound. Rather than assuming that a feature is present or absent in PAS, it is necessary to describe a feature as a function of its degree of presence in PAS. This assumption is similar to Chomsky and Halle's (1968) distinction between the classificatory and phonetic function of distinctive features. The features are assumed to be binary in their classificatory function, but not in their phonetic or descriptive function. In the latter, features are multivalued representations that describe aspects of the speech sounds in the perceptual representation. Similarly, Ladefoged (1975) has also distinguished between the phonetic and phonemic level of feature description. A feature describing the phonetic quality of a sound has a value along a continuous scale whereas a feature classifying the phonemic composition is given a discrete value. In our framework, the continuous features in PAS are transformed into discrete percepts in synthesized auditory memory (SAM) by the primary recognition process.

Given this theoretical description, acoustic features in PAS must be expressed as continuous values. That is to say, the listener

will be able to hear the degree of presence or absence of a particular feature, even though his judgment in a forced choice task will be discrete. Oden and Massaro (1977) have used this description to describe acoustic features as fuzzy; that is to say, varying continuously from one speech sound to another. In this representation features are represented as fuzzy predicates which may be more or less true rather than only absolutely true or false (Zadeh, 1971). In terms of the model, fuzzy predicates represent the feature detection and evaluation process; each predicate is applied to the speech sound and specifies the degree to which it is true that the sound has a relevant acoustic feature. For example, rather than assuming that a sound is voiced or voiceless, the voicing feature of a sound is expressed as a fuzzy predicate.

$$P(\text{voiced}(S_{ij})) = .65 \qquad (1)$$

The predicate given by Equation 1 represents the fact that it is .65 true that speech sound S_{ij} is perceived to be voiced. In terms of our model, then, the feature detection process makes available a set of fuzzy predicates at the level of PAS. In addition to being concerned with the acoustic features in preperceptual storage this analysis of the feature evaluation process makes apparent that an important question in speech perception research is how the various continuous features are integrated into a synthesized percept.

As an example of the study of acoustic features, consider the dimension of voicing of speech sounds. In English, the stops, fricatives, and affricates can be grouped into cognate pairs that

have the same place and manner of articulation but contrast in voicing. The question of interest is what acoustic features are responsible for this distinction and how the various features are integrated together in order to provide the perceptual distinction. The integration question has not been extensively studied, however, since the common procedure in these experiments is to study just a single acoustic feature at a time. Consider two possible cues to the voicing distinction in stop consonant syllables: voice onset time (VOT), the time between the onset of the syllable and the onset of vocal cord vibration, and the fundamental frequency (F_0) of vocal cord vibration at its onset. Each of these cues have been shown to be functional in psychophysical experiments when all other cues have been held constant at neutral values. But it is difficult to generalize these results to the perception of real speech, since no information is provided about the weight that these features will carry when other features are also present in the signal. To overcome this problem, it is necessary to independently vary two or more acoustic features in the signal. The results of this type of experiment will not only provide information about the cue value of one feature when other features are present in the signal, but will also allow the investigator to evaluate how the various acoustic features are combined into an integrated percept. (For a further discussion see Massaro & Cohen, 1976, in press; Oden & Massaro, 1977).

b. Acoustic features in fluent speech

The success of finding acoustic features in perception of isolated speech sounds might lead one to expect that perception of fluent speech is a straightforward process. Sound segments could be recognized on the basis of their features and the successive segments could be combined into higher-order units of words, phrases, and sentences. However, the acoustic structure of words in fluent speech differ significantly from the same words spoken in isolation. Two souces contribute to the large variation of the words in fluent speech: coarticulation and psychological parsimony (Cole, & Jakimik, 1977; Ross, 1975). In fluent speech, the speech articulators must assume an ordered series of postures corresponding to the intended sounds, and the articulators cannot always reach their intended targets because of influence of adjacent movements. Coarticulation refers to altering the articulation of one sound because of neighboring sounds. The words did and you spoken as /did/ and /ju/ in isolation will be articulated as /dIdʒu/ in combination because of palatalization. The alveolar stop followed by a front glide when combined produce the front-palatal affricate /dʒ/, even though a word boundary intervenes. Psychological parsimony, sometimes called laziness (Ross, 1975), refers to the minimization of effort when we speak (Lieberman, 1967; Ross, 1975). Extending our example, did you can be further modified to give /dIdʒə/ or just /dʒə/ in the utterance Did you want to go? Therefore, we get the message when a close friend asks /dʒə wanə go/ or even /jə wanə go/?

Luckily, the speaker is not only lazy but also intelligent. He anticipates the linguistic competence of his audience and the contextual constraints in the message (Lieberman, 1967). For example, a speaker will usually tend to give the listener a better acoustic signal for words that have high information content. Lieberman (1963) asked listeners to identify words excised from continuous speech. Identification was good to the degree that the excised word was unpredictable in the original utterance. The word "nine" was recognized about twice as often when it was excised from the sentence, "The number you will hear is nine," than when it was taken from "A stitch in time saves nine." If a word is not highly predictable from context, the speaker compensates by providing the listener with a better acoustic signal. In a heroic study, Umeda (1977) measured the temporal properties of consonant sounds in 20 minutes of speech. Content words had longer durations than function words, and she interprets these results in terms of the high information value of content relative to function words.

3. Primary Recognition

The primary recognition process evaluates the acoustic features in PAS and compares or matches these features against those defining perceptual units in long-term memory (LTM). Every perceptual unit has a representation in long-term memory, which is called a sign or prototype. The prototype of a perceptual unit is specified in terms of the acoustic features that define the ideal acoustic information

as it would be represented in PAS. The recognition process operates to find the prototype in LTM which best matches the acoustic features in PAS. The outcome of this process is the phenomenalogical experience of hearing a particular sound at some location in space. This synthesized percept is held in synthesized auditory memory (SAM). In contrast to the feature detection process, the outcome of the primary recognition process is influenced by the listener's knowledge and expectations and can be modified by learning experience. Two issues are critical for understanding the nature of the primary recognition process. The first issue is the properties of preperceptual auditory storage (PAS) and the second is the nature of perceptual units.

a. Preperceptual auditory storage

Preperceptual auditory storage holds the features passed on by the detection process for a short time after a sound is presented. Given that a speech sound is temporally extended, its acoustic features can be detected at varying times during or after the speech sound. Furthermore, different features might require different amounts of time for feature detection. Given that the features do not enter preperceptual storage simultaneously, they must be integrated across some short period. In speech perception, preperceptual auditory storage accumulates the acoustic features of a speech stimulus until the sound pattern is complete. Primary recognition occurs during and possibly after this time in order to arrive at a synthesized auditory percept. A second pattern does not

usually occur until the first pattern has been perceived. However, if the second pattern is presented soon enough, it should interfere with recognition of the first pattern. By varying the delay of the second pattern, we can determine the duration of preperceptual auditory storage and the temporal course of the recognition process. The experimental task is referred to as a backward recognition masking paradigm (Massaro, 1972).

In one study (Massaro, 1974) the consonant-vowel (CV) syllables /ba/, /da/, and /ga/ were used as test and masking stimuli. Only enough of each syllable was presented to make it sound speech-like. The 42-msec syllables had 30 msec of CV transition plus 12 msec of steady state vowel. On each trial, 1 of the 3 syllables was presented followed by a variable silent interval before presentation of a second syllable chosen from the same set of 3 syllables. The subject's task was to identify the first syllable as one of the 3 alternatives, and to ignore the second syllable, if possible. The speech sounds were presented at a normal listening intensity.

Figure 2 plots the observed results in terms of discriminability (d') values for each of the three test alternatives. The d' measure of signal detection theory provides an index of how well the subject discriminates a given test alternative from the other test alternatives in the task. This measure allows the experimenter to correct for any decision bias the subject may have under a particular experimental condition. The d' values are computed from the hit and false alarm probabilities. The probability of identifying a syllable

correctly is designated as a hit, and responding with that syllable to any other syllable alternative is designated a false alarm. For example, the probability of responding /ba/ given the test alternative /ba/, P(ba|ba), would be a hit, whereas the probability of responding /ba/ to /da/ or /ga/, P(ba|da or ga) would be a false alarm. The d' value given by these two independent probabilities indexes the discriminability of the syllable /ba/.

Figure 2 shows that correct identification of the first speech sound increased with increases in the silent interval between the two sounds. These results show that recognition of the consonant phoneme was not complete at the end of the CV transition or even at the end of the short vowel segment of the sound. Syllable recognition required perceptual processing after the speech sound was terminated. The second speech sound interfered with perception if it was presented before recognition was complete. These results support the idea that the speech sound is held in preperceptual auditory storage while processing takes place. A second sound interferes with any further resolution of the first sound.

What do these results imply about the on-line processing of continuous speech? We will argue below that V, CV, and VC syllables function as perceptual units in speech processing. The backward masking experiment with CV syllables shows that the consonant is not recognized before the vowel, but rather the CV syllable is recognized as a unit. In our model, accurate recognition requires sufficient processing time after the information

of the sound pattern is placed in PAS. Massaro (1972, 1974) has presented some evidence that V and CV syllables are processed during the steady-state vowel period. In this view, the extended vowel periods in continuous speech are redundant in terms of providing additional segmental information, but could serve the important function of allowing for sufficient processing time before new information is presented. Silent periods also allow time for processing and it has been shown that silent time after a VC syllalbe is critical for accurate recognition (Abbs, 1971). (For a further discussion of processing time in continuous speech, see Massaro, 1975b).

b. Perceptual units in speech

In the framework of the information processing model, the primary recognition process integrates the featural information held in preperceptual storage into a percept in synthesized memory. One question relevant to the model is the functional units at this stage of processing. The functional units are called perceptual units which correspond to units that are described in long-term memory. The primary recognition process finds the best match between the featural information in PAS and the descriptions of perceptual units in long-term memory. The recognition process, then, involves the transformation of the featural information into a synthesized percept.

We have argued that perceptual units correspond to sound patterns of V, CV, or VC size since these units can be described by relatively invariant acoustic features (Massaro, 1975b). Smaller

units such as phonemes lack invariance, and in fact, this lack of invariance has been one of the central foci of speech perception theory (cf. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Perceptual units of larger size have also been proposed. However, the results in backward recognition masking shows that preperceptual auditory storage cannot maintain featural information for a period of time greater than roughly one quarter of a second. Therefore, some transformation must take place on the order of every syllable; phrases, clauses, or even words are inappropriate perceptual units at the primary recognition stage of processing. (For further discussion, see Massaro, 1975b,Chapter 4.)

4. Secondary Recognition

Secondary recognition transforms synthesized percepts into meaningful forms in generated abstract memory. In speech perception, it is assumed that the input is analyzed syllable by syllable for meaning. The secondary recognition process makes the transformation from percept to meaning by finding the best match between the perceptual information in SAM and the lexicon in long-term memory. Each word in the lexicon contains both perceptual and conceptual information. The concept recognized is a function of two independent sources of information: the perceptual information in synthesized memory and the syntactic/semantic context in the message. It should be noted that the latter source of information does not have to be in the message per se but could result from the situational context and the knowledge of the listener.

a. Perceptual and contextual contributions

We assume that the secondary recognition process operates syllable by syllable on the output of primary recognition. In this sense, our conceptualization of speech processing is one that is perceptually, and, therefore, acoustically driven. However, contextual constraints also exert a strong influence at this stage of processing, so that both contributions must be accounted for in describing how meaning is imposed on the spoken message. A series of recent studies have shown that abstracting meaning is a joint function of the perceptual and contextual information. In one experiment, Cole (1973) asked subjects to push a button every time they heard a mispronunciation in a spoken rendering of Lewis Carroll's Through the Looking Glass. A mispronunciation involved changing a phoneme by 1, 2, or 4 distinctive features (for example, confusion mispronounced as gunfusion, bunfusion, and sunfusion, respectively). The probability of recognizing a one-feature mispronunciation was .3 whereas a four-feature change was recognized with probability .75. This result makes apparent the contribution of the perceptual information passed on by the primary recognition process. Some of the mispronunciations went unnoticed because in our view the contribution of contextual information worked against the recognition of a mispronunciation. The syntactic/semantic context of the story would support a correct rendering of the mispronounced word, outweighing the perceptual information. In support of this idea, all mispronunciations were

correctly recognized when the syllables were isolated and removed from the passage.

A second paradigm that has been used to study speech processing is the shadowing task, in which the listener repeats back the message as it is heard. It is well-known that shadowing performance improves with increases in the syntactic/semantic constraints in the message (Rosenberg & Lambert, 1974; Treisman, 1965). Recent research has been directed at how these higher-order constraints are integrated with the ongoing perceptual analyses in order to arrive at the meaning of the message. Marslen-Wilson (1973) asked subjects to shadow prose as quickly as they heard it. Some individuals were able to shadow the speech at extremely close delays with lags of 250 msec, about the duration of a syllable or so. When subjects made errors in shadowing, the errors were syntactically and semantically appropriate given the preceding context. For example, given the sentence "He had heard at the Brigade," some subjects repeated "He had heard that the Brigade." In this example, that shares acoustic information with at and is also syntactically/semantically appropriate in the same position in the sentence.

In another experiment (Marslen-Wilson, 1975), subjects shadowed sentences that had one of the syllables mispronounced in a three-syllable word. Subjects never restored the word, that is, repeated back what should have been said when the mispronunciations occurred in the first syllable. With mispronunciations in the second and third syllables, however, a significant proportion of restorations

occurred. If the mispronounced word was syntactically and semantically anomalous, however, restorations did not occur for any mispronounced syllable. These results indicate that restorations will not occur if the shadower does not have sufficient acoustic information and syntactic/semantic context to make the restoration appropriate. If context were the exclusive and overriding factor, we might expect subjects to replace the syntactically-semantically anomalous word with the appropriate word. This did not occur, however, showing that both context and acoustic information influenced speech processing.

Marslen-Wilson and Tyler (1975) had subjects monitor sentences for a target item in three types of target monitoring tasks. The target item was either a particular word, any word that rhymed with the target, or a member of superordinate category. Three types of sentences were used: (1) normal, (2) syntactically correct but semantically anomalous by randomization of the content words, and (3) a completely random ordering of the words in the sentence. The mean reaction times for detecting the target was a function of both the monitoring task and the sentence structure. The reaction times were shortest for detecting a specific target, next shortest for a rhyme, and longest for a member of a superordinate category. Sentence structure facilitated monitoring in all three tasks, however, showing that higher-order constraints were functional, regardless of the nature of the target analyses that were required.

Marslen-Wilson and Welsh (in press) asked observers to shadow (repeat back) spoken passages from a popular novel. The words of the passage were read to the subjects at a rate of 160 words per minute. The subjects were told to repeat back exactly what they heard. At random throughout the passage, common three-syllable words were mispronounced. When the words were mispronounced, only a single consonant phoneme was changed to a new consonant phoneme. The new phoneme differed from the original by one or three phonemic distinctive features, based on Keyser and Halle's (1968) classification system. Independently of the degree of feature change the changes could occur in the first or third syllable of the three-syllable word. Finally, the mispronounced words were either highly predictable or unpredictable given the preceding portion of the passage. Subjects were not told that words could be mispronounced although they probably became aware of this early in the experiment. All subjects shadowed at relatively long delays greater than 600 msec. The primary dependent measure in the task was the percentage of fluent restorations, that is, the proportion of times the shadowers repeated what should have been said rather than what was said. About half of the mispronounced words were restored and the restorations were made on-line with an average latency, and the shadowing was not disrupted. (When the mispronunciation was repeated exactly, i.e., not restored, shadowing was disrupted and response times increased.)

The change in the percentage of restorations as a function of the 3 independent variables in Marslen-Wilson and Welsh's study can

illuminate how acoustic information and high-order context are integrated by the listener in language processing. Figure 3 presents the observed results in terms of the percentage of fluent restorations. All three variables influenced the likelihood of a restoration: shadowers were more likely to restore a one-feature than a three-feature change, a change in the third than in the first syllable, and a change in a highly predictable than in an unpredictable word.

Marslen-Wilson and his colleagues interpret this series of experiments as evidence against serial theories of language processing, which assume that "varying degrees of delay before information at any one level of analysis can interact with information at a higher level" (Marslen-Wilson & Tyler, 1975, p. 784). However, the results do show exactly such a delay. Restorations seldom occur when the first syllable is mispronounced by 3 features even though the word is relatively probable given the preceding context. This means that some low-level perceptual analyses of the word occurred regardless of the higher-order constraints available and then the outcome of these analyses was combined with the higher-order constraints. The fact that higher-order constraints in the passage influence shadowing does not mean that some analyses do not begin before others. More importantly, their view might be interpreted to mean that higher-order analyses modify the output of lower-level analyses. However, a quantitative model that assumes that both levels of analyses are functionally independent

can accurately describe the results of their experiments. Figure 3 presents the predictions of a quantitative formulation of the independence model. The model assumes that the information passed on by the feature detection and evaluation process is equivalent regardless of the higher order constraints in the message. Therefore, it is not necessary to assume that higher-order constraints allow the subject to selectively attend to or selectively process certain acoustic properties of the speech input. In this sense, we argue that higher-order constraints do not modify the low-level nature of the perceptual analyses performed on the input data.

5. Rehearsal and Recoding

In the present model, the same abstract structure stores the meaning of both listening and reading. Generated abstract memory (GAM) in our model corresponds to the working memory of contemporary information processing theory. Rehearsal and recoding processes operate at this stage to maintain and build semantic/syntactic structures. There is good evidence that this memory has a limited capacity, holding about 5 ± 2 chunks of information. For a more detailed discussion of processing at this stage, see Massaro (1975a, Chapter 27). Rehearsal and Recoding operations are the workhorses of the simultaneous translation task and it is at this stage the task becomes unique relative to normal language processing.

a. Generated abstract memory

Although GAM is assumed to be abstract relative to SAM and SVM, the nature of the information appears to be tied to the surface structure of the language rather than in terms of underlying meaning that is language independent. Relevant research comes from work experiments carried out with bilingual subjects (Dornic, 1975, provides an excellent review). Recall from immediate memory (supposedly tapping GAM) does not differ for unilingual and bilingual lists, whereas recall of items assumed to be no longer in GAM was poorer in bilingual than unilingual lists (Tulving & Colotla, 1970). Similarly, Kintsch and Kintsch (1969) showed that the semantic relationship between the words in different languages did not influence immediate memory, but did affect recall of items no longer active in GAM. Salgert, Hamayan and Ahmar (1975) showed that multilingual subjects remembered the specific language of words in a mixed language list of unrelated words, but this information was forgotten when the words were presented in sentence contexts. Dornic (1975) points out that surface structure and item information are integrally related in immediate memory; subjects seldom report translations for the words. If the items are remembered, so are the appropriate surface structure forms.

b. Rehearsal

Although some work has been carried out on rehearsal and recoding operations, it is not clear how relevant it is to the simultaneous translation situation. In our model, GAM has a "limited capacity"

and the learning and memory for information is a direct function of rehearsal and recoding processes. Memory of an item will increase with the time spent operating on that item, and will decrease with the time spent operating on other "unrelated" items. This "limited capacity" rule has provided a reasonable description of the acquisition and forgetting of information in GAM (cf. Massaro, 1975a, Chapter 27).

c. Recoding

The simultaneous translator must recode the surface code of the source language to that of the target language. A critical question at this stage of processing centers around the size of the units that are recoded. It seems unlikely that recoding occurs word by word given that many words are ambiguous until later context disambiguates their meaning. It might be assumed, therefore, that the translator builds syntactic/semantic structures in the source language and then translates these structures into abstract forms and finally recodes the abstract structures into the target language.

The success of the simultaneous translator depends, in part, on how quickly the recognition of the original units of the message can be recoded into their appropriate transformations. The degree of stimulus-response compatibility may be important in this regard. Consider an experiment carried out by Alluisi, Strain, and Thurmond (1964). Subjects were visually presented with a single digit under

three different levels of stimulus uncertainty. Subjects had to either name the digit, name the digit that was one larger than the test digit, or name another number that had been previously assigned to the digit at random. Reaction times increased with increases in stimulus uncertainty on all response conditions, but at a much faster rate for incompatible than for compatible responses. Responding with a low-compatible response will require much more time than with a high-compatible response, especially when there is high stimulus uncertainty.

Theios (1975) has formulated a model in which each name code in a stimulus is associated with a hierarchy of responses. Selecting a response high in the hierarchy is fast and relatively independent of stimulus uncertainty. To the extent the appropriate response is low in the hierarchy, response selection time will be slow and depend on stimulus uncertainty. In this view, the simultaneous interpreter must organize his response hierarchy so that the surface forms of the target message are high and other forms are low.

6. Long-Term Memory

The structure of long-term memory in terms of the representation of meaning has been a central concern of "cognitive scientists" during the last decade (cf. Kintsch & Wilkes, /this volume). Lexical storage is usually assumed to be a necessary and central component of the representation of meaning. We view the subjective lexicon as a multi-dimensional representation with both perceptual and conceptual attributes. The perceptual codes of wind consist of the sound of

the spoken word wind, the look of the letters that spell wind, a picture of a windy scene, and the sound of the wind blowing, and so on. The conceptual code consists of the relatively abstract (but fuzzy) properties that define the meaning of wind, such as air movement. Language understanding involves going from perceptual codes to conceptual ones. whereas production goes in the reverse direction. Secondary recognition performs this function in understanding, whereas recoding must be involved in production.

The translator must have perceptual codes of both languages stored with the conceptual codes. Also, the two sets of perceptual codes must be differentiated at some level, so that memory access and retrieval will be limited to the appropriate surface form. One central issue in discussions of the memory structures of multilinguals is whether the two languages share the same memory or whether each language has a unique and separate memory (cf. Dornic, 1975). In terms of lexical representations, would there be additional codes for wind in the second language or would there be another unique entry for wind in the second language. In the former case, additional perceptual codes are added to the same conceptual codes, whereas in the latter, new conceptual codes would be established along with the new perceptual codes of the second language. As Dornic (1975) points out, however, neither of these extremes may be correct with the truth lying somewhere in between.

If bilinguals have a single conceptual code, then accessing this code makes available more perceptual codes than would be available for

the unilingual. Traditional results and theory in reaction-time research would then lead to the conclusion that response times should be longer for bilinguals; reaction-time usually increases with increases in the number of response alternatives (Ervin, 1961). Given that bilinguals do not seem to show this deficit, they must be able to filter out one language and switch in the other with very little decrement in performance. Dornic (1975) has provided some evidence for this capability by showing that a set and expectation for one language interferes with processing a second language. Many of the issues in bilingual research seem to be basic to understanding simultaneous interpretation.

7. Conclusion

It is hoped that analyzing simultaneous interpretation in terms of an information-processing framework is more than an academic exercise. Many of the skills required in translation are currently being studied in language-processing research. Knowledge acquired in this work should be directly applicable to understanding and training interpreters. Gerver (1976) has admirably reviewed empirical studies of simultaneous interpretation and has outlined a model similar to the one proposed here. The permanent structural features and control processes in his model are similar to the structural and functional components in the present model. Finally, it should be mentioned that Moser has had some success in extending and exploiting these models in the development of training programs for interpreters. I look forward to further developments that will provide the critical question for information-processing theory: how well does it work in the real world?

References

Abbs, M. H. A study of cues for the identification of voiced stop consonants in intervocalic contexts. Unpublished dissertation, University of Wisconsin, 1971.

Alluisi, E. A., Strain, G. S., & Thurmond, J. B. Stimulus response compatibility and the rate of gain of information. Psychonomic Science, 1964, 1, 111-112.

Bernstein, L. The unanswered question. The 1973 Norton Lectures at Harvard. Cambridge, Mass.: Harvard Univ. Press, 1976.

Chomsky, N., & Halle, M. The sound pattern of English. New York: Harper & Row, 1968.

Cole, R. A. Listening for mispronunciations: A measure of what we hear during speech. Perception & Psychophysics, 1973, 13, 153-156.

Cole, R. A., & Jakimik, J. Understanding speech: How words are heard. Technical Report, Department of Psychology, Carnegie-Mellon University, 1977.

Dornic, S. Human information processing and bilingualism. Report from the Institute of Applied Psychology. The University of Stockholm, No. 67, 1975.

Ervin, S. M. Semantic shift in bilingualism. American Journal of Psychology, 1961, 74, 233-241.

Jakobson, R., Fant, C. G. M., & Halle, M. Preliminaries to speech analysis: The distinctive features and their correlates. Cambridge, Mass.: MIT Press, 1961.

Keyser, S. J., & Halle, M. What we do when we speak. In P. A. Kolers and M. Eden (Eds.), Recognizing patterns. Cambridge, Mass.: MIT Press,1968.

Kintsch, W., & Kintsch, E. Interlingual inference and memory processes. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 16-19.

Gerver, D. Empirical studies of simultaneous interpretation: A review and a model. In R. Brislin (Ed.) Translation: Applications and research. New York: Gardner Press, 1976.

Ladefoged, P. A course in phonetics. New York: Harcourt, Brace, and Jovanovich, 1975.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code, 1967, 74, 431-461.

Lieberman, P. Some effects of semantic and grammatical context on the production and perception of speech. Language & Speech, 1963, 6, 172-187.

Lieberman, P. Intonation, perception, and language. Cambridge, Mass.: MIT Press, 1967.

Marslen-Wilson, W. Linguistic structure and speech shadowing at very short latencies. Nature, 1973, 244, 522-523.

Marslen-Wilson, W. D. Sentence perception as an interactive parallel process. Science, 1975, 189, 226-228.

Marslen-Wilson, W., & Tyler, L. K. Processing structure of sentence perception. Nature, 1975, 257,784-786.

Marslen-Wilson, W., & Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, in press.

Massaro, D. W. Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review, 1972, 79, 124-145. (a)

Massaro, D. W. Perceptual units in speech recognition. Journal of Experimental Psychology, 1974, 102, 199-208.

Massaro, D. W. Understanding language: An information-processing model of speech perception, reading, and psycholinguistics. New York: Academic Press, 1975. (b)

Massaro, D. W. Experimental psychology and information processing. Chicago: Rand-McNally, 1975. (a)

Massaro, D. W., & Cohen, M. J. The contribution of fundamental frequency and voice onset time to the /zi/ and /si/ distinction. Journal of the Acoustical Society of America,1976, 60, 704-717.

Massaro, D. W., & Cohen, M.M. Voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. Perception & Psychophysics, in press.

Oden, G. C., & Massaro, D. W. Integration of place and voicing information in identifying synthetic stop-consonant syllables. WHIPP Report #1, Wisconsin Human Information Processing Program, July, 1977.

Rosenberg, S., & Lambert, W. E. Contextual constraints and the perception of speech. Journal of Experimental Psychology, 1974, 102, 178-180.

Ross, J. R. Parallels in phonological and semantic organization. In J. F. Kavanagh & J. E. Cutting (Eds.), The role of speech in language. Cambridge, Mass.: MIT Press, 1975.

Theios, J. The components of response latency in human information processing. In S. Dornic and P.M.A. Rabbitt (Eds.), Attention and Performance V, New York: Academic Press, 1975.

Treisman, A. M. Verbal responses and contextual constraints in language. Journal of Verbal Learning and Verbal Behavior, 1965, 4, 118-128.

Tulving, E., & Colotla, V. A. Free recall of trilingual lists. Cognitive Psychology, 1970, 1, 86-98.

Umeda, N. Consonant duration in American English. Journal of the Acoustical Society of America, 1977, 61, 846-858.

Zadeh, L. A. Quantitative fuzzy semantics. Information Sciences, 1971, 3, 159-176.

List of Figures

1. Flow diagram of temporal course of understanding a language pattern.
2. Accuracy of recognition of the three voiced stop-consonants (measured in d' values) as a function of the silent interval between the offset of the test syllable and the onset of the masking syllable. The points give the observed results and the lines give the predictions of a quantification of the primary recognition process.
3. The percentage of fluent restorations of mispronunciations in a shadowing task as a function of the degree of feature change of the mispronounced sounds, whether the mispronuncation occurred in the first or third syllable of the word, and whether the mispronounced word was highly or lowly predictable given the preceding context. (Observed results from Marslen-Wilson & Welsh, in press.)

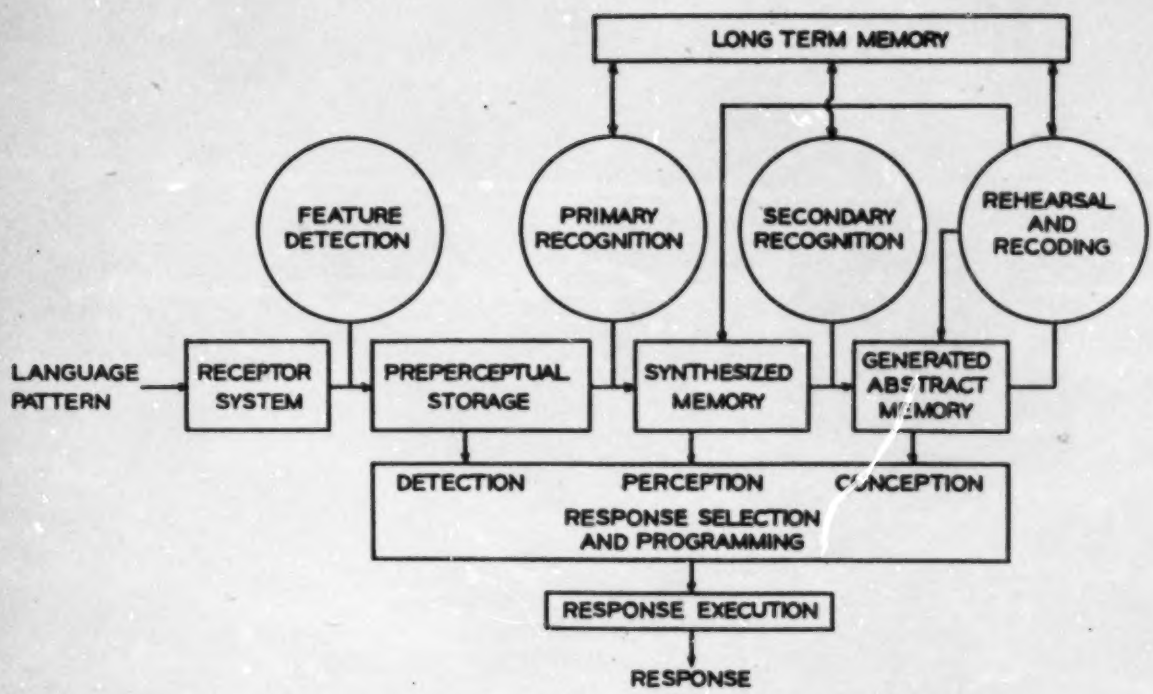
LONG TERM MEMORY
FEATURE DETECTION
PRIMARY RECOGNITION
SECONDARY RECOGNITION
REHEARSAL AND RECODING
LANGUAGE PATTERN
RECEPTOR SYSTEM
PREPERCEPTUAL STORAGE
SYNTHESIZED MEMORY
GENERATED ABSTRACT MEMORY
DETECTION
PERCEPTION
CONCEPTION
RESPONSE SELECTION AND PROGRAMMING
RESPONSE EXECUTION
RESPONSE

Fig 1

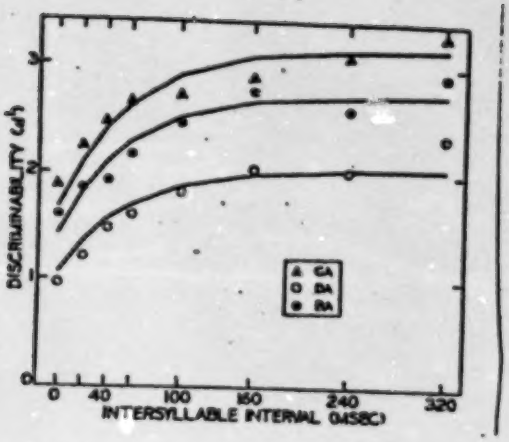
DISCRIMINABILITY (d')
0
1
2
3
0
40
100
150
240
320
INTERSYLLABLE INTERVAL (MSEC)
CA
BA
PA

Fig 2

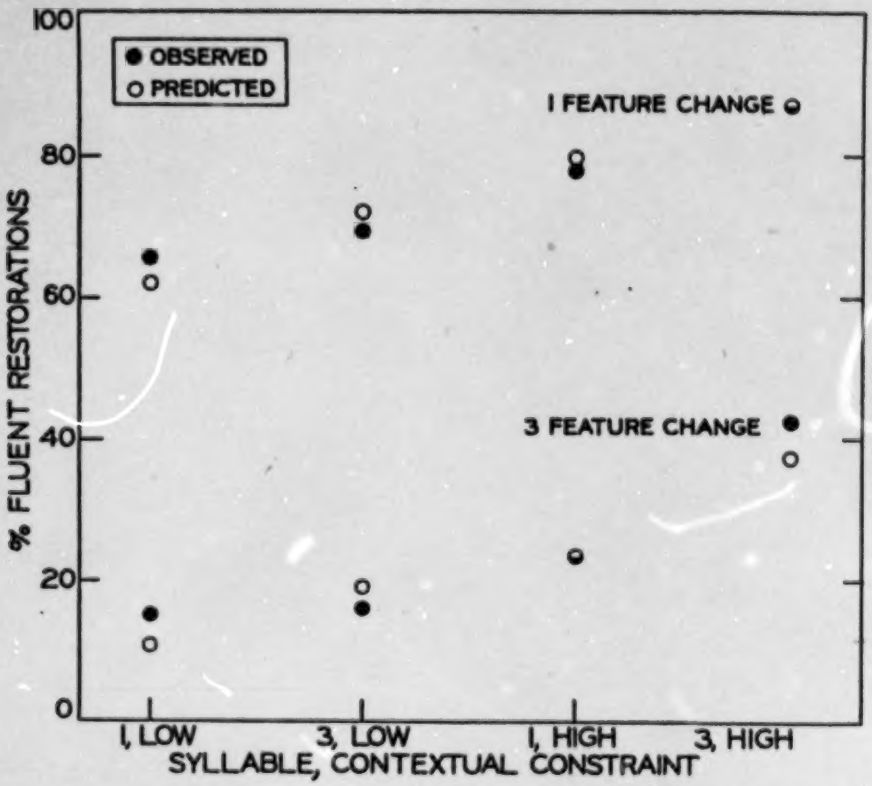
100
80
60
40
20
0
% FLUENT RESTORATIONS
● OBSERVED
○ PREDICTED
1 FEATURE CHANGE
3 FEATURE CHANGE
1, LOW
3, LOW
1, HIGH
3, HIGH
SYLLABLE, CONTEXTUAL CONSTRAINT

Fig 3